

学号姓名_课程研学报告（二）

学号：

姓名：

日期：

课程研学报告 撰写格式要求： 正文宋体 5 号，A4，1.25 行距

课程报告撰写内容：

一、构建查询需求

依据“课程研学报告（一）”中得到的查询 q 构建多个查询需求，如 query : **Research on Development Architecture of WeChat Platform**，可以构建查询需求如下所示。

Q1 : WeChat Platform

Q2 : Development WeChat

Q3 : Development WeChat Platform

Q4 : Architecture WeChat

Q5 : Architecture WeChat Platform

Q6 : Development Architecture WeChat

Q7 : Development Architecture WeChat Platform

二、构建理想相关文档集

将“课程研学报告（一）”中计算得到的、与 q 的相似度最高的前 10 篇文档作为理想相关文档集 Rel 。

三、选择搜索系统

在“百度学术”、“Bing 学术”和“Google 学术”中，选择一个作为搜索系统。

四、评价查询 Q 的检索效果

在所选搜索系统中，对每个查询需求进行检索，所得到有序检索结果的前 50-100 条记录，作为查询结果集 A 。如 $Q1 \sim Q7$ 得到的查询结果集为 $A1 \sim A7$ 。

PRECISION *Precision (P) is the fraction of retrieved documents that are relevant*

$$(8.1) \quad \text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} = P(\text{relevant}|\text{retrieved})$$

RECALL *Recall (R) is the fraction of relevant documents that are retrieved*

$$(8.2) \quad \text{Recall} = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})} = P(\text{retrieved}|\text{relevant})$$

These notions can be made clear by examining the following contingency table:

(8.3)

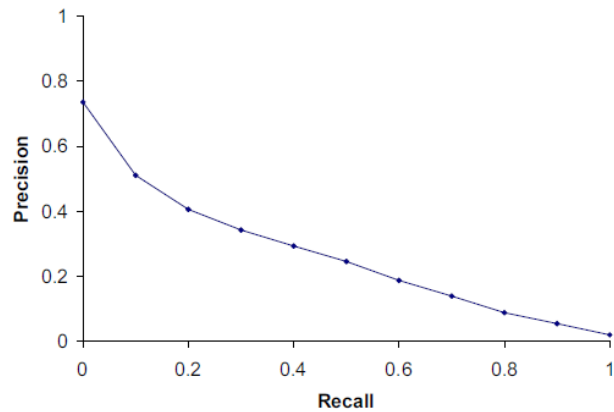
	Relevant	Nonrelevant
Retrieved	true positives (tp)	false positives (fp)
Not retrieved	false negatives (fn)	true negatives (tn)

Then:

$$(8.4) \quad \begin{aligned} P &= tp / (tp + fp) \\ R &= tp / (tp + fn) \end{aligned}$$

【说明】: $|Rel| = tp + fn$ $|A| = tp + fp$

利用（复习课件）例题的计算方法，采用公式 8-4 计算**每个**查询结果集的 **Precision** 和 **Recall**，并画出 **P-R 曲线**。



利用公式（如下图所示），计算查询 **q** 的平均 **Precision**，并画出平均 **P-R 曲线**。

Average Precision (平均查准率)

$$\bar{P}(r) = \sum_{i=1}^{N_q} \frac{P_i(r)}{N_q}$$

N_q 是使用的查询总数

$P_i(r)$ 是查全率为 r 时，第 i 个查询的查准率

评价方法：

- 对每个查全率下的查准率进行平均化->平均查准率